Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios

Dapozo, Gladys; Porcel, Eduardo; López, María V.; Bogado, Verónica
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura
Universidad Nacional del Nordeste. 9 de Julio Nº 1449. CP 3400. Corrientes. Argentina.
TE: (03783) 423126 - (03783) 473930 Fax
{gndapozo, eporcel, mvlopez}@exa.unne.edu.ar; vro.s.bg@gmail.com

RESUMEN

El estudio del gran volumen de información que se obtiene de los alumnos que ingresan a la universidad permitirá lograr una caracterización de los mismos. Esto servirá de punto de partida para relacionar estos datos con otras variables que contribuyan a identificar situaciones o factores que estén relacionados con el bajo rendimiento académico de los estudiantes en el primer año de carrera universitaria. Para este propósito, se han aplicado técnicas de minería de datos mediante una herramienta de software libre. Sin embargo, a través de las pruebas realizadas, se ha detectado una gran cantidad de datos inconsistentes, incoherentes y, principalmente, faltantes. Debido a esto, se propone analizar, en profundidad, las técnicas de preprocesamiento de datos disponibles en los programas que implementan técnicas de *datamining* con el fin de incrementar la calidad de los datos previo a su procesamiento, principalmente, y por otra parte, detectar posibles dificultades de interpretación de los requerimientos del formulario diseñado para recabar la información por parte de los aspirantes a ingresar a la universidad.

Palabras clave: Minería de datos. Técnicas de preprocesado de datos. Herramienta de software libre. Rendimiento académico de alumnos universitarios.

INTRODUCCIÓN

La preocupación por el desempeño de los alumnos de primer año de carrera universitaria, que surge de los desfavorables indicadores de desgranamiento, abandono y rendimiento académico, ha llevado a las universidades del país a investigar sobre las causas que subyacen en esta problemática. La Universidad Nacional del Nordeste (UNNE) no es ajena a esta situación. En este sentido, ha realizado varios estudios con el objeto de aportar información que contribuya a configurar un cuadro de situación al interior de la institución [1] [2]. La Facultad de Ciencias Exactas de la UNNE, con una matrícula de más de 6.000 alumnos, de los cuales el 30% corresponde a la Licenciatura en Sistemas de Información, no escapa a la realidad descripta anteriormente. En esta Facultad, el Grupo de investigación en Matemática Aplicada a la Investigación Educativa, ha venido realizando desde el año 2000, trabajos de investigación que se han publicado en distintas revistas científicas y presentado, entre otros congresos, en las Reuniones Científicas que cada año organiza la Universidad , y que pueden consultarse en su página Web[1]. Como subproyecto de este proyecto macro, se pretende abordar el análisis de los datos a través de técnicas de minería de datos. Un trabajo previo de los autores [3], dejó al descubierto las debilidades en cuanto a la calidad y completitud de los datos y la necesidad de estudiar también, en función de los resultados, la eficiencia del formulario diseñado para recabar los datos de los ingresantes.

La minería de datos puede ofrecer para esta problemática planteada una gran variedad de métodos estadísticos y computacionales para investigar la existencia de relaciones y patrones de comportamiento en datos almacenados electrónicamente. Estas relaciones o patrones emergentes pueden sugerir explicaciones causales que puedan ser verificadas posteriormente o bien pueden sugerir estrategias de acción para lograr ciertos objetivos de cambio [4].

[1] http://www.unne.edu.ar/Web/cyt/presentacion.php

Antes de aplicar cualquier técnica de minería de datos es preciso realizar un análisis previo de los datos de que se dispone.
La primera tarea que se suele abordar es el análisis exploratorio y gráfico de los datos. La mayoría del software estadístico dispone de herramientas que aportan técnicas gráficas preparadas para el examen de los datos que se ven mejoradas con medidas estadísticas más detalladas para su descripción. Estas técnicas permiten el examen de las características de la distribución de las variables implicadas en el análisis [5].
La segunda tarea es el análisis de los datos ausentes. Cualquier recogida y proceso de datos presenta problemas que van a impedir obtener información de algunos de los elementos de la población en estudio, como las negativas a colaborar, las ausencias de los encuestados en el momento de la toma de datos, la inaccesibilidad de algunos elementos o los errores en los instrumentos de medida. El analista deberá identificar la presencia de datos ausentes y llevar a cabo las acciones necesarias para intentar minimizar sus efectos [5].
Las posibles soluciones para el problema de los valores faltantes son: a) Ignorar: Algunos algoritmos son robustos a datos faltantes (ej: árboles de decisión); b) Eliminar/Reemplazar toda la columna: Si hay muchos valores faltantes no nos servirá. A veces se puede "rehacer" a partir de otra/s columnas dependientes; c) Eliminar la Fila: Si tenemos muchas instancias con valores faltantes, nos quedamos sin ejemplos; d) Reemplazar el Valor por la media, varianza o moda o bien predecirlo [6].
La tercera tarea es la detección de valores atípicos, que también suelen denominarse *outliers*. Se trata de detectar la existencia de observaciones que no siguen el mismo comportamiento que el resto. Los casos atípicos suelen deberse a errores en el procedimiento a la hora de introducir los datos o de codificarlos. Una vez detectados los casos atípicos el analista debe saber elegir con criterio entre eliminarlos del análisis o evaluar toda la información incluyéndolos [5].
Existe una primera categoría de datos atípicos formada por aquellas observaciones que provienen de un error de procedimiento, por ejemplo un error de codificación, error de entrada de datos, etc. Estos datos atípicos, si no se detectan mediante filtrado, deben eliminarse o recodificarse como datos ausentes. Una segunda categoría de casos atípicos contempla aquellas observaciones que ocurren como consecuencia de un acontecimiento extraordinario existiendo una explicación para su presencia en la muestra. Éstos generalmente se retienen en la muestra, salvo que su significancia no sea relevante. Una tercera categoría de datos atípicos comprende las observaciones extraordinarias para las que el investigador no tiene explicación, las cuales normalmente se eliminan del análisis. Una cuarta categoría de casos atípicos la forman las observaciones que se sitúan fuera del rango ordinario de valores de la variable. Suelen denominarse valores extremos y se eliminan del análisis si se observa que no son elementos significativos para la población. Las propias características del caso atípico, así como los objetivos del análisis que se realiza, determinan los casos atípicos a eliminar [5].
Para detectar valores atípicos resulta útil hacer uso de los histogramas. La forma de detectarlos y las medidas a tomar dependen mucho del dominio.
Las posibles soluciones al problema de los valores erróneos son: a) Ignorar: Algunos algoritmos son robustos; b) Eliminar la columna; c) Eliminar la fila; d) Reemplazar el valor por: nulo, máximo o mínimo; d) Discretizar y hacer que los anómalos sean "muy alto" o "muy bajo" [6].
Por otra parte, siguiendo esta línea de trabajo, se ha diseñado e implementado un *datawarehouse* (almacén de datos) que integra toda la información sistematizada de los alumnos de la Facultad de Ciencias Exactas y Naturales y Agrimensura (FACENA) de la UNNE. El mismo contiene los datos de todas las actividades académicas de los alumnos, como asignaturas cursadas y rendidas, trámites de reinscripción y readmisión, reconocimiento de materias y datos del egreso o trámite de graduación [7]. Esta base de datos permitirá realizar diferentes análisis orientados al seguimiento de los alumnos en función de esta caracterización inicial.

El objetivo de este trabajo es lograr, a través de técnicas de preprocesamiento de datos, mejorar la calidad de los datos correspondientes a los alumnos ingresantes a la FACENA-UNNE, utilizando herramientas de software libre.

METODOLOGIA

La metodología dentro de la fase de preparación de los datos no está estandarizada, existen varias versiones según los autores. Para este trabajo se seleccionó la descripta en [9] que está compuesta por tres etapas principales, las cuales se dividen en otras subetapas de propósito específico:

Etapa 1. Selección de datos

Se utilizarán los datos de los alumnos provenientes del denominado Sistema de Ingreso de Alumnos de la UNNE. Este sistema incluye un formulario, en el cual los aspirantes a ingresar a la universidad hacen constar, además de sus datos de identificación personal, sus principales antecedentes sociodemográficos tales como edad, sexo, estado civil, lugar de procedencia y de residencia y sus antecedentes educacionales tales como tipo de título y de colegio de nivel medio del cual provienen, así como el nivel educativo alcanzado por los padres y el tipo de trabajo y categoría ocupacional de los mismos. Este formulario es único para toda la universidad, se completa y registra en la base de datos de cada Facultad y se centraliza posteriormente, al cierre del período de inscripción, en la base de datos de la Secretaría de Planeamiento de la UNNE, la cual distribuye luego estos datos a los investigadores que los requieran.

Un punto importante relacionado con esta información tiene que ver con la modalidad de registro de los datos a través del formulario a llenar por los alumnos, quienes en general completan la ficha sin ningún tipo de asesoramiento especializado, y esta situación deja librada a su comprensión las consignas ó códigos establecidos en dicho formulario.

Etapa 2. Preprocesado de Datos

- **Integración de Datos:** Los datos, si bien corresponden a la misma información, presentan formatos diferentes en los diferentes períodos de tiempo, debido a modificaciones del instrumento de recolección (formulario), de manera que deberán ser sometidos a un proceso de integración y unificación de conceptos.
- **Reconocimiento y Limpieza de Datos:** Este paso tiene como objetivo reducir el ruido y las inconsistencias. Para ello, se seleccionará un resumen (muestra) de los datos, pudiendo de esta forma interpretar la validez de algún valor para algún atributo y mejorar la calidad de los datos. En el presente trabajo resulta interesante estudiar la cantidad de datos nulos y los *outliers* (datos erróneos), con el fin de reducir las posibilidades de error cuando sean procesados por el algoritmo de minería. Se aplicarán y analizarán las técnicas que facilitan el reconocimiento y limpieza de los datos que provee la herramienta seleccionada (Weka), siendo algunas de éstas las que se explican brevemente a continuación.
  - Tablas de Resumen de Atributos: En Weka denominada Relación Actual, es una relación (tabla) que contiene todos los datos necesarios, considerados relevantes para el estudio.
  - Resúmenes de Estadísticas: La herramienta proporciona un resumen de las estadísticas para el atributo que se está evaluando: porcentaje de instancias con valores nulos, cantidad de diferentes valores para el atributo, cantidad de instancias que tienen un valor único para ese atributo (diferente de las demás instancias).
  - Diagramas: Se emplea un diagrama de barras segmentadas para poder visualizar cada valor distinto del atributo y la cantidad de instancias que presentan este valor.

Etapa 3. Selección de Características

- **Transformación de Datos:** Consiste en la normalización de los mismos. Este paso implica la transformación del tipo de algunos atributos, en caso que fuera necesario, teniendo presente que convertir el tipo de un atributo a otro puede cambiar la semántica de dicho atributo. Este paso está muy ligado al algoritmo que procesará los datos para obtener conocimiento. Algunas

técnicas comúnmente usadas son: discretización y escalado y centrado (estandarización). La técnica a usar en este paso para el presente trabajo quedará sujeta a la técnica de minería seleccionada posteriormente.

- **Reducción de Datos:** En este paso se disminuye el tamaño de los datos, eliminando características redundantes.
  - Selección/Extracción de Atributos: Existen varias técnicas para llevar a cabo la tarea de selección de los atributos relevantes como ser: Métodos Basados en Filtros, mediante los cuales se filtran los datos antes de ser procesados por el algoritmo, y Métodos Basados en Wrappers, que seleccionan los atributos en función de la calidad del modelo de minería asociado a los atributos utilizados.
    La extracción de atributos puede ser vista como una proyección del espacio de estudio, ya que permite transformar el espacio de atributos, obteniendo otro espacio de atributos que represente la misma información de diferente manera.
  - Construcción de Atributos: Si se presentan patrones complejos en los datos se construirá un atributo sencillo de interpretar por el algoritmo. Para realizar esta tarea se pueden recurrir a diferentes técnicas: construcción guiada por los datos, por el modelo ó por el conocimiento.

Esta última etapa se realizará cuando se haya analizado y seleccionado algunas de las técnicas de extracción de conocimiento que se adapten al lote de datos, ya que ésta puede que requiera la aplicación de alguna técnica particular para cada subetapa mencionada.
Las técnicas de preprocesamiento de datos a utilizar son las que provee la herramienta Weka (*Waikato Environment for Knowledge Analysis*) de la Universidad de Waikato, software que se encuentra de manera gratuita en el sitio oficial de esta institución en Internet y contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas [8].

OBJETIVOS PROPUESTOS

Este subproyecto tiene los siguientes objetivos:

- Incrementar la calidad de los datos de los alumnos de la FACENA correspondientes a los años de ingreso 2000 al 2005.
- Mejorar los resultados descriptivos en cuanto a la caracterización de los ingresantes respecto de datos generales: género, edad, estado civil, situación laboral, lugar de procedencia, nivel educativo previo y nivel socioeconómico y nivel educativo de sus padres.
- Seleccionar las variables del formulario de ingreso, que resulten relevantes para ser incluidas en modelos matemáticos que permitan analizar su influencia sobre otras variables (extraídas de otros estudios dentro del contexto del proyecto macro), tales como los hábitos de estudio y los conocimientos matemáticos previos de los ingresantes.
- Incorporar la información depurada de los ingresantes de la FACENA al *datawarehouse* que contiene toda la información sistematizada de las actividades de los alumnos regulares de la Facultad, especialmente adaptada para los propósitos de estudio de la problemática de los factores asociados con el bajo rendimiento académico de los alumnos. Esta incorporación permitirá ampliar el espectro de análisis y cruce de variables.
- Informar, si fuera necesario, a las autoridades competentes de la UNNE, acerca de la eficiencia del formulario diseñado para la recolección de datos de los ingresantes, a fin de mejorar la comprensión de las consignas por parte de los alumnos.

CONCLUSIONES

Los resultados que se obtengan de este subproyecto permitirán, por un lado, contribuir a brindar más información respecto de la problemática objetivo del proyecto en cuyo contexto se actúa. Esta información podrá orientar decisiones o acciones concretas destinadas a mejorar los preocupantes índices de desgranamiento, abandono y bajo rendimiento de los alumnos en el primer año de universidad.

Por otra parte, se pretende incrementar el conocimiento sobre las distintas técnicas de preprocesamiento de datos, dada la importancia que tiene esta etapa en la aplicación de minería de datos o en cualquier otro tipo de análisis de información, como así también, verificar las posibilidades reales de las herramientas de software libre para el análisis de grandes volúmenes de datos.

Finalmente, otro propósito fundamental está relacionado con la formación de recursos humanos. Se trata de consolidar el equipo de trabajo en esta temática de la Informática y de la Estadística, para avanzar posteriormente hacia aplicaciones de minería de datos en el área de las empresas privadas o estatales que se caracterizan por un importante grado de informatización y que, en general, no aprovechan en su totalidad el gran volumen de datos existentes, para obtener mayor conocimiento de sus actividades que permitan orientar en forma más eficiente las acciones en pro de sus objetivos institucionales.

REFERENCIAS

[1]Ojeda, Gabriel Eduardo. "La Secretaría General de Planeamiento y la Investigación educativa" en http://eluniversitario.unne.edu.ar/2004/44/pagina/01informeespecial02.htm visualizado el 14/02/2007.

[2] Foio, Socorro. "El perfil socioeconómico de los ingresantes en la UNNE y su relación con la deserción en el primer año, la retención y el rendimiento académico" en http://www.unne.edu.ar/Web/estadistica/temainteres/Texto/Inf_Ingres/inf_ingres.htm visualizado el 14/02/2007.

[3] Dapozo, G., Porcel, E., López, M., Bogado,V., Bargiela, R., "Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE". Anales del VIII Workshop de Investigadores en Ciencias de la Computación (WICC) ISBN 950-9474-35-5.Universidad de Morón. 2006.

[4] Sananes, Marta; Torres, Elizabeth; Sinha, Surendra P. y Nava Puente, Luis. "Búsqueda y caracterización de subgrupos de pobreza mediante la aplicación de algunas técnicas de minería de datos". Instituto de Estadística Aplicada y Computación, Universidad de Los Andes, Mérida, Venezuela. Escuela de Estadística, Universidad de Los Andes, Mérida, Venezuela.

[5] Pérez López, C. "Técnicas de análisis multivariante de datos. Aplicaciones con SPSS". Editorial Pearson Prentice-Hall. Madrid. España . pp 21-22, 39-40, 48. 2004.

[6] Cortizo Pérez, J. C. "Preprocesado de datos". D. Sistemas Informáticos. Esc. Superior Politécnica. Universidad Europea de Madrid. AINetLab (AINetSolutions). http://www.ainetsolutions.com

[7] Dapozo, G., Porcel, E. "Metodología de integración de datos para apoyar el seguimiento y análisis del rendimiento académico de los alumnos de la FACENA". Comunicaciones Científicas y Tecnológicas de la UNNE 2005. http://www.unne.edu.ar/Web/cyt/com2005/8-Exactas/E-032.pdf.

[8] Machine Learning Project at the Department of Computer Science of The University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/

[9] Gustavo González Sánchez, Sonia Delfín Ávila, Josep Lluís de la Rosa. Preprocesamiento de bases de datos masivas y multi-dimensionales en minería de uso web para modelar usuarios: comparación de herramientas y técnicas con un caso de estudio Institut d'Informàtica i Aplicacions, Agents Research Lab, Universidad de Girona, España.
http://eia.udg.es/~gustavog/esp/publicaciones/cedi2005_gustavo_sonia_published.pdf